CLAIMS

1. A method operative in a distributed network having a set of servers each provisioned with a manager process and an application framework on which application components are executed in response to service requests, comprising:

generating a first data set identifying which application components are actually loaded on which servers;

publishing the first data set to each manager process in the set;

responsive to a service request received at a manager process on a given server, determining whether an instance of a given application component is loaded on the given server; and

if the instance of the given application component is not loaded on the given server, directing the service request to another server in the set of servers as a function of information in the first data set.

2. The method as described in claim 1 wherein the service request is directed to a server that, as indicated by the first data set, the given application component is loaded.

3. The method as described in claim 2 further including the step of:

generating a second data set identifying which application components should be loaded on which servers.

4. The method as described in claim 3 wherein the service request is directed to a server on which the given application component is loaded irrespective of whether the given application component, as indicated by the second data set, should be loaded on the server.

5. The method as described in claim 3 further including, at a given server, the step of loading one or more application components on the given server according to

information in the second data set.

6. The method as described in claim 3 further including, at a given server, the step of unloading one or more application components from the given server according to information in the second data set.

7. A method operative in a distributed network having a set of server region, each server region having a set of servers each provisioned with a manager process and an application framework on which application components are executed in response to service requests, comprising:

generating a first data set identifying, for a given server region, which application components are actually loaded on which servers;

generating a second data set identifying, for the given server region, which application components should be loaded on which servers;

publishing the first and second data sets to each manager process in the set of servers in the given server region;

responsive to a service request received at a manager process on a given server in the given server region, determining whether an instance of a given application component is loaded on the given server; and

if the instance of the given application component is not loaded on the given server, directing the service request to another server in the set of servers in the given server region as a function of information in the first and second data sets.

8. The method as described in claim 7 wherein the service request is directed to a server in the given server region that, as indicated by the second data set, the given application component should be loaded and that, as indicated by the first data set, the given application component is actually loaded.

9. The method as described in claim 7 wherein the service request is directed to a server in the given server region that, as indicated by the first data set, the given application component is actually loaded irrespective of whether the second data set indicates that the application component should be loaded.

10. The method as described in claim 7 further including the step of including, at a given server, loading one or more application components on the given server

according to information in the second data set.

11. The method as described in claim 7 further including, at a given server, the step of unloading one or more application components from the given server according to information in the second data set.

12. Apparatus for use with a set of servers in a distributed network, each server having a manager process for managing service requests, an application framework on which application components are executed in response to the service requests, and a set of one or more application components, comprising:

code for generating a first data set identifying which application components are actually loaded on which servers within the set;

code for generating a second data set identifying which application components should be loaded on which servers within the set; and

code for balancing load across the set of servers based on information in the first and second data sets.

13. The apparatus as described in claim 12 wherein the code for balancing load implements a first policy with respect to service requests that cannot be serviced by a given manager process at a given server.

14. The apparatus as described in claim 13 wherein the first policy directs a service request from the given server to another server that, as indicated by information in the first data set, the given application component is loaded.

15. The apparatus as described in claim 13 wherein the first policy directs a service request from the given server to another server that, as indicated by the second data set, the given application component should be loaded and that, as indicated by the first data set, the given application component is actually loaded.

16. The apparatus as described in claim 12 wherein the code for balancing load implements a second policy with respect to one or more application components on a given server.

17. The apparatus as described in claim 16 wherein the second policy loads a

given application component on the given server according to information in the second data set.

18. The apparatus as described in claim 16 wherein the second policy unloads a given application component from the given server according to information in the second data set.